Lecture 24

(b) Non equal covariance metrics $\Sigma_1 \neq \Sigma_2$

(1) Quatratic rule based on MECM

- Population case

Allocate X_0 to

π_1 if $-\frac{1}{2}X_0'(\Sigma_1^{-1}-\Sigma_2^{-1})X_0 + (\mu_1'\Sigma_1^{-1}-\mu_2'\Sigma_2^{-1})X_0 - K \geq \ln \gamma$

π_2 if - - - - - - - - $< \ln \gamma$

Where $K = \frac{1}{2}\ln\frac{|\Sigma_1|}{|\Sigma_2|} + \frac{1}{2}(\mu_1'\Sigma_1^{-1}\mu_1 - \mu_2'\Sigma_2^{-1}\mu_2)$

- Sample case

1°) $\hat{\mu}_i = \bar{X}_i$, $\hat{\Sigma}_i = S_i$

2°) Allocate X_0 to

π_1 if $-\frac{1}{2}X_0'(S_1^{-1}-S_2^{-1})X_0 + (\bar{X}_1'S_1^{-1} - \bar{X}_2'S_2^{-1})X_0 - \hat{K} \geq \ln \gamma$

if ———————— $< \ln \gamma$

Where $K = \frac{1}{2}\ln\frac{|S_1|}{|S_2|} + \frac{1}{2}(\bar{X}_1'S_1^{-1}\bar{X}_1 - \bar{X}_2 S_2^{-1}\bar{X}_2)$

(2) Practical issues

- check normality first
- If necessary, employ transformation method to get more normal data
- test equality of $\Sigma_1 = \Sigma_2$ (Chapter 6.6) Likelihood ratio test (chi-square test)
- Try different classification procedures and compare their performance and select the best.

11.4 Evaluating classification procedures

(a) Optimal error rate (misclassification probability)

(1) Total probability of misclassification (TPM)

$$TPM = P_1 P(2|1) + P_2 P(1|2)$$
$$= P_1 \int_{R_2} f_1(x)\,dx + \int_{R_1} P_2 f_2(x)\,dx$$
$$= P_1 + \int_{R_1} P_2 f_2(x) - P_1 f_1(x)\,dx$$

(2) Optimal error rate (OER)

- To minimize TPM we should select

$$R_1 : \{x : P_2 f_2(x) - P_1 f_1(x) \le 0\}$$
$$\Leftrightarrow \{x : \frac{f_1(x)}{f_2(x)} \ge \frac{P_2}{P_1}\}$$

- $$OER = P_1 + \int_{\{x: \frac{f_1(x)}{f_2(x)} \ge \frac{P_2}{P_1}\}} (P_2 f_2(x) - P_1 f_1(x))\,dx$$

minimized TPM

(3) Actual error rate (AER)

- In practice, after we have $\hat{R}_1$ and $\hat{R}_2$ We obtain an estimate of OER, called actual error rate (AER)

$$AER = P_1 \int_{\hat{R}_2} f_1(x)\,dx + P_2 \int_{\hat{R}_1} f_1(x)\,dx$$

- AER still depends on $f_i(x)$, p_i, $i=1,2$

(b) Apparant error rate (APER)

(1) APER

- Applicable for any classification rules
- It is the fraction of observations in the "training" sample that are misclassified
- Confusion matrix predicted membership

Actual membership	π_1	π_2	
π_1	n_{1C}	n_{1M}	n_1
π_2	n_{2M}	n_{2C}	n_2

- $$APER = \frac{n_{1M} + n_{2M}}{n_1 + n_2}$$

(2) Advantages and disadvantages

- Easy to calculate (popular)
- But usually can underestimate the AER (unless n_1 and n_2 are very large)
- Reuse of data (building and judging)

(C) Some other procedure and expected actual error rate

- Spliting sample

1°) Split sample into

{ training sample → to construct classification rule
{ validation sample (test) → to check performance

2°) Disadvantages

requires large sample

may lose valuable information

- Cross-validation to estimate expected AER

(1) cross-validation (jackknife, leave-one-out, hold out)

1°) start with π_1, hold out one point from π_1

develop a classification rule based on $n_1 - 1$, n_2

2°) classify the "hold out" point unsing the rule in 1°)

3°) repeat 1°) and 2°) untill all points in π_1 are classified

Let $n_{1M}^{(H)}$ be the # of misclassified points in π_1

4°) repeat 1°) - 3°) for π_2, obtain $n_{2M}^{(H)}$

(2) Estimate of expected AER

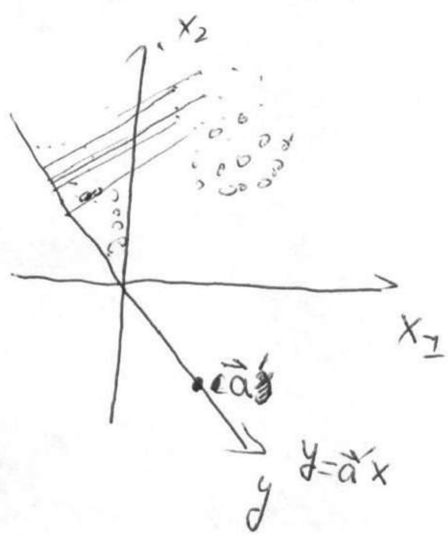
$$\hat{E}(AER) = \frac{n_{1M}^{(H)} + n_{2M}^{(H)}}{n_1 + n_2}$$

11.5 Fisher's Discriminant function — Seperation of two populations ($\Sigma_1 = \Sigma_2$) (No normal assumption)

(a) basic idea and classification rule

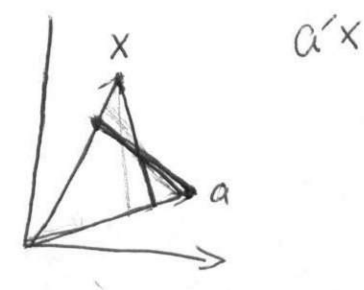
(1) Multivariate to univariate via transformation

- Transfom the multivariate observation X from π_1 and π_2 to univariate y such that the y's derived from π_1 and π_2 are seperated as fas as possible



$\vec{a} = (a_1, a_2)'$
$X = (X_1, X_2)'$

$a'x$



1°) Let $X_{11}, \cdots, X_{1n_1}$ and $X_{21}, \cdots X_{2n_2}$ are samples from π_1 and π_2

Let a be a vector, project X_{ij} to a, $i = 1, 2$, $j = 1 \cdots n$

Let $y = a'x$ then

$y_{11} = a'X_{11}$	$y_{21} = a'X_{21}$
$\vdots$	$\vdots$
$y_{1n_1} = a'X_{1n_1}$	$y_{2n_2} = a'X_{2n_2}$
π_1	π_2
$\bar{y}_1 = a'\bar{X}_1$	$\bar{y}_2 = a'\bar{X}_2$

2°) Let $S_{iy}^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 = a' S_{ix} a$

$$S_y^2 = \frac{(n_1-1) S_{1y}^2 + (n_2-1) S_{2y}^2}{n_1+n_2-2} = a' S_{pooled}\, a$$

$$S_{pooled} = \frac{(n_1-1) S_{1x} + (n_2-1) S_{2x}}{n_1+n_2-2}$$

Define seperation $= \frac{\bar{y}_1 - \bar{y}_2}{s_y}$

3°) Find $\hat{a}$ that maximizes

$$\text{Seperation}^2 = \frac{(\bar{y}_1 - \bar{y}_2)^2}{s_y^2} = \frac{(a'(\bar{X}_1 - \bar{X}_2))^2}{a' S_{pooled}\, a}$$

4° Since $\frac{(a'(\bar{x}_1 - \bar{x}_2))^2}{a' S_{pooled}\, a}$

$$= \frac{\left((S_{pooled}^{\frac{1}{2}} a)' S_{pooled}^{-\frac{1}{2}} (\bar{X}_1 - \bar{X}_2)\right)^2}{(S_{pooled}^{\frac{1}{2}} a)' (S_{pooled}^{\frac{1}{2}} a)}$$

$$= \frac{(u'v)^2}{\|u\|^2} \leq \frac{(\|u\| \|v\|)^2}{\|u\|^2} = \|v\|^2$$

$$= (\bar{X}_1 - \bar{X}_2)' S_{pooled}^{-\frac{1}{2}} S_{pooled}^{-\frac{1}{2}} (\bar{X}_1 - \bar{X}_2)$$

The equality is satisfied only if

$u = cv$ Let $c = 1$

$$\Rightarrow S_{pooled}^{\frac{1}{2}} a = S_{pooled}^{-\frac{1}{2}} (\bar{X}_1 - \bar{X}_2)$$

$$\Rightarrow \hat{a} = S_{pooled}^{-1} (\bar{X}_1 - \bar{X}_2)$$

(2) Fisher's discrimination fuction (classification) rule

- $\hat{a}'X$ is called Fisher's discriminant function

Let $\hat{y} = \hat{a}'x$, then allocate x_0 to

$$\begin{cases} \pi_1 : \text{if } \hat{y}_0 = \hat{a}'x_0 \geq \frac{1}{2}(\bar{y}_1 + \bar{y}_2) = \hat{a}'\left(\frac{\bar{x}_1 + \bar{x}_2}{2}\right) \\ \pi_2, \text{if } \cdots < \cdots \end{cases}$$

(equivalent to MECM rule for $r=1$, under normal populations with $\Sigma_1 = \Sigma_2 = \Sigma$)

(b) Test for $\mu_1 = \mu_2$ (under normal assumption)

(1) Maximum seperation and the test (See chapter 6)

- Consider $H_0: \mu_1 = \mu_2$

If $\pi_i : N_p(\mu_i, \Sigma)$ then

$$\left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{-1} D^2 \sim \frac{(n_1+n_2-2)P}{n_1+n_2-P-1} F_{P, n_1+n_2-P-1}$$

$$D^2 = (\bar{X}_1 - \bar{X}_2)' S_{pooled}^{-1} (\bar{X}_1 - \bar{X}_2)$$

- reject H_0 if D^2 is large (not imply Good classification)

- If seperation is not significant, the search for a useful classification rule will probably prove fruitless.

11.6 classification with several population (g populations, $g \geq 2$)

(a) MECM method and rule

<1> the method

- P_i ($i=1,\cdots,g$) prior probability

  $P(k|i) = P(X \text{ allocated to } \pi_k \mid x \in \pi_i)$

  $= \int_{R_k} f_i(x)\,dx \qquad i, k = 1, \cdots g$

  R_k: the set of x's classified as π_k

  $C(k|i)$ = the cost of allocates x to π_k while it belongs to π_i

- $ECM(1) = \sum_{k=1}^{g} C(k|1)\,P(k|1)$

  $= \sum_{k \neq 1} C(k|1)\,P(k|1)$

  $ECM(i) = \sum_{k \neq i} C(k|i)\,P(k|i)$

  $ECM = \sum_{i=1}^{g} P_i\,ECM(i)$

  $= \sum_{i=1}^{g} P_i \left(\sum_{k \neq i} C(k|i)\,P(k|i) \right)$

(2) The rule

- Allocate x to π_K if

$$\sum_{\substack{i=1 \\ i \neq K}}^{g} p_i f_i(x) c(K|i) \leq \sum_{\substack{i=1 \\ i \neq j}}^{g} p_i f_i(x) c(j|i) \quad \text{fall all } j=1,2,\cdots g \text{ and } j \neq K$$

- If all misclassification costs are equal, then allocate x to π_K if

$$p_K f_K(x) \geq p_i f_i(x) \quad \text{for all } i \neq K$$

$$\Updownarrow$$

$$p(\pi_K|x) = \frac{p_K f_K(x)}{\sum_j p_j f_j(x)} \geq \frac{p_i f_i(x)}{\sum_j p_j f_j(x)} = p(\pi_i|x) \quad \text{for all } i \neq K$$

"posterior probability"

$$\Updownarrow$$

$$\ln p_K f_K(x) \geq \ln p_i f_i(x) \quad \text{for all } i \neq K$$

(b) classify normal populations

<1> Unequal Σ_i

- Quadratic discriminant Score

1°) Population Case.

Consider π_i : $N_p(\mu_i, \Sigma_i)$, $i=1,\cdots,g$. then

$$\ln p_i f_i(x) = \ln p_i + \ln f_i(X)$$
$$= \ln p_i - \frac{p}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma_i| - \frac{1}{2}(X-\mu_i)'\Sigma_i^{-1}(X-\mu_i)$$

according to the MECM rule with $c(k|i)$ all the same,

Allocate x to π_k if

$$\ln p_k f_k(x) = \max_i \ln p_i f_i(x)$$

Define the quadratic discriminant score as

$$d_i^Q(x) = -\frac{1}{2}\ln|\Sigma_i| - \frac{1}{2}(x-\mu_i)'\Sigma_i^{-1}(x-\mu_i) + \ln p_i$$

2°) Sample case

$$\hat{d}_i^Q(x) = -\frac{1}{2}\ln|S_i| - \frac{1}{2}(x-\bar{x}_i)'S_i^{-1}(x-\bar{x}_i) + \ln p_i$$

- MECM rule (with equal misclassification costs)

2°) Sample case

$$x \to \pi_k \quad \text{if } \hat{d}_k(x) = \max_{1\le i\le g} \hat{d}_i^Q(x)$$

1°) population case

$$x \to \pi_k \quad \text{if } d_k^Q(x) = \max_{1\le i\le g} d_i^Q(x)$$

(2) Equal Σ_i linear

- linear discriminant ~~case~~ score

1°) population case

$$\ln p_i f_i(x) = \ln p_i - \left(\frac{p}{2}\right)\ln 2\pi - \frac{1}{2}\ln|\Sigma| - \frac{1}{2}(x-\mu_i)'\Sigma^{-1}(x-\mu_i)$$

$$= \boxed{\ln p_i - \left(\frac{p}{2}\right)\ln} - \left(\frac{p}{2}\right)\ln 2\pi - \frac{1}{2}\ln|\Sigma| - \frac{1}{2}x'\Sigma^{-1}x$$

$$+ \mu_i'\Sigma^{-1}x - \frac{1}{2}\mu_i'\Sigma^{-1}\mu_i + \ln p_i$$

Define the linear discriminant score as

$$d_i(x) = (\Sigma^{-1}\mu_i)' x - \frac{1}{2}\mu_i'\Sigma^{-1}\mu_i + \ln p_i, \quad i = 1, \cdots, g$$

2°) Sample case

$$\hat{d}_i(x) = (S_{pooled}^{-1}\,\bar{x})' x - \frac{1}{2}(\bar{x}_i)' S_{pooled}\, \bar{x}_i + \ln p_i$$

$$S_{pooled} = \frac{(n_1 - 1)S_1 + \cdots + (n_g - 1)S_g}{(n_1 + \cdots + n_g) - g}$$

- MECM with equal $c(k|i)$'s

1°) population case

$$x \to \pi_k \quad \text{if } d_k(x) = \max_i d_i(x)$$

2°) sample case

$$x \to \pi_k \quad \text{if } \hat{d}_k(x) = \max_i \hat{d}_i(x)$$

$$\Leftrightarrow \quad -\frac{1}{2}\underbrace{(x - \bar{x}_k)' S_{pooled}^{-1} (x - \bar{x}_k)}_{D_k^2} + \ln p_k$$

$$= \max_i \left(-\frac{1}{2} D_i^2 + \ln p_i\right)$$

(c) Examples

11.7 Fisher's discriminant functions for seperation of sereral populations

(a) Idea and classification rule

(1) Idea and Assumption

- Idea: Find a few linear combinations

$$a_1'X,\quad a_2'X,\ \ldots$$

Which could be used to represent populations and seperate populations as much as possible

- Assumption $\Sigma_1 = \Sigma_2 = \cdots = \Sigma_g = \Sigma$

(2) population case

- Let μ_i be the mean of π_i and $\bar{\mu} = \frac{1}{g}\sum_{i=1}^{g}\mu_i$
- Let $B_\mu = \sum_{i=1}^{g}(\mu_i - \bar{\mu})(\mu_i - \bar{\mu})'$ measures between group variability

  between groups sum of cross products
- Consider $Y = a'X$ then

$$Var(Y) = a'V(X)\,a = a'\Sigma a = \sigma_Y^2$$

$$E(Y) = a'E(X) = a'\mu_i \quad \text{if } X \in \pi_i$$
$$= \mu_{iY}$$

Let $\bar{\mu}_Y = \frac{1}{g}\sum_{i=1}^{g}\mu_{iY} = \frac{1}{g}\sum_{i=1}^{g}a'\mu_i = a'\bar{\mu}$

- Define

$$\text{Seperation}^2 = \frac{\sum_{i=1}^{g}(\mu_{iY} - \bar{\mu}_Y)^2}{\sigma_Y^2}$$

$\leftarrow$ sum of squared distances from populations to overall mean of Y

$$= \frac{a' B_\mu a}{a' \Sigma a}$$

$\leftarrow$ The common variability within groups

- To find a to maximize seperation^2

(3) Sample case

- Let $X_{i1}, \dots, X_{in_i}$ be a sample from π_i

Let $\bar{X}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij} \longrightarrow$ estimate μ_i

$\bar{X} = \frac{\sum_{i=1}^{g} n_i \bar{X}_i}{\sum_{i=1}^{g} n_i} \longrightarrow$ estimate $\bar{\mu}$

$S_{pooled} = \frac{\sum_{i=1}^{g}(n_i-1)S_i}{\sum_{i=1}^{g} n_i - g} = \frac{W}{\sum_{i=1}^{g} n_i - g} \longrightarrow$ estimate Σ

(*) $B = \sum_{i=1}^{g} n_i (\bar{X}_i - \bar{X})(\bar{X}_i - \bar{X})' \longrightarrow B_\mu$

- Find $\hat{a}$ to minimize

$$\frac{a'Ba}{a'S_{pooled}a} \Leftrightarrow \frac{a'Ba}{a'Wa} = \frac{(W^{\frac{1}{2}}a)'W^{-\frac{1}{2}}BW^{-\frac{1}{2}}(W^{\frac{1}{2}}a)}{(W^{\frac{1}{2}}a)'(W^{\frac{1}{2}}a)}$$

$$= \frac{u'W^{-\frac{1}{2}}BW^{-\frac{1}{2}}u}{u'u} = \frac{u'}{\|u\|}W^{-\frac{1}{2}}BW^{-\frac{1}{2}}\frac{u}{\|u\|} \leq \lambda_{max}(W^{-\frac{1}{2}}BW^{-\frac{1}{2}})$$

$$= \lambda_{max}(W^{-1}B)$$

$$W^{\frac{1}{2}}\hat{a} = e_1(W^{-\frac{1}{2}}BW^{-\frac{1}{2}})$$

$$W^{-\frac{1}{2}}BW^{-\frac{1}{2}} \cdot W^{\frac{1}{2}}\hat{a} = \lambda_{max} \cdot W^{\frac{1}{2}}\hat{a}$$

$$W^{-1}B\hat{a} = \lambda_{max}\hat{a}$$

$$\Rightarrow \quad \hat{a} = e_1(W^{-1}B)$$

(4) Fisher's sample linear discriminants

- Let $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_s > 0$, $S \leq \min(g-1, p)$ be nonzero eigenvalues of $W^{-1}B$ and $\hat{e}_1, \cdots, \hat{e}_s$ be the corresponding eigen vectors

Then $\hat{a}_1 = \hat{e}_1$ maximizes

$$\frac{a'Ba}{a'Wa} \quad \text{or} \quad \frac{a'Ba}{a'S_{pooled}a}$$

- $\hat{y}_1 = \hat{a}_1'x = \hat{e}_1'x$ is called the sample first discriminant

$\hat{y}_2 = \hat{a}_2'x = \hat{e}_2'x$ ———— second ————

$\vdots$

$\hat{y}_k = \hat{a}_k'x = \hat{e}_k'x$ ———— kth ————

$\vdots$

(5) Fisher's classification procedure based on sample discriminant

- Let $\hat{y}_j = \hat{a}_j' x$, $\bar{y}_{kj} = \hat{a}_j' \bar{x}_k$ $\quad j=1,\cdots,r$, $r \le s$

1°) Allocate $x \to \pi_k$ if

$$\sum_{j=1}^{r} (\hat{y}_j - \bar{y}_{kj})^2 = \sum_{j=1}^{r} [\hat{a}_j'(x - \bar{x}_k)]^2$$

$$\le \sum_{j=1}^{r} (\hat{y}_j - \bar{y}_{ij})^2 = \sum_{j=1}^{r} [\hat{a}_j'(x - \bar{x}_i)]^2, \quad \forall i \ne k$$

2° Let $\hat{y} = \begin{pmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_r \end{pmatrix}$, $\bar{y}_k = \begin{pmatrix} \bar{y}_{k1} \\ \vdots \\ \bar{y}_{kr} \end{pmatrix} = \begin{pmatrix} \hat{a}_1' \bar{x}_k \\ \vdots \\ \hat{a}_r' \bar{x}_k \end{pmatrix}$

$$= \begin{pmatrix} \hat{a}_1' x \\ \vdots \\ \hat{a}_r' x \end{pmatrix}$$

Then allocate $x \to \pi_k$ if

$$\|\hat{y} - \bar{y}_k\|^2 \le \|\hat{y} - \hat{y}_i\|^2, \quad \forall i \ne k$$

```
options ls=85 ps=65;
title1 'SAS DISCRIM example 2';
title2 h=1 'Admission data for graduate school of business; JW: T11-6';
data admission;
 infile 'Z:\\My Documents\Teaching\Stat524\Fall 2010\Data set\T11-6.dat' firstobs=1;
 input gpa gmat status$;
  gpa1=200*gpa;
run;

proc print data=admission;
run;

proc plot data=admission;
   plot gpa*gmat=status;
run;

title2 "scatter plot of gpa and gmat";
%plotit(data=admission,labelvar=_blank_, symvar=status,
plotvars=gpa1 gmat, color=black, colors=blue);
run;

proc discrim data=admission method=normal pool=test
     wcov pcov manova listerr crosslisterr;
 class status;
 var gpa gmat;
*priors equal/propotional/'1'=0.3 '2'=0.5 '3'=0.1;
run;

proc discrim data=admission method=normal pool=yes wcov pcov
      listerr crosslisterr;
 class status;
 var gpa gmat;
run;

proc discrim data=admission method=normal pool=no wcov pcov
      listerr crosslisterr;
 class status;
 var gpa gmat;
run;
```

Lecture 26

Chapter 12 Cluster analysis

<1> Introduction

- clustering and classification

(1) Classification, Gives the number of groups, assigns new observation to one of the groups.

(2) clustering : (grouping): No assumption on the number of groups or the structure of groups. searches for some "natural" groups of items (varibles) based on similarities (or dissimilarities)

12.2 Similarity measures

(a) two types of problems in cluster analysis

(1) Cluster obs (items, units, cases) based on some sort of "distance" (numerical measurements)

(2) cluster variables based on correlation coefficients

(b) Similarity measurements for pairs of items

(1) Euclidean distance

$$X = \begin{pmatrix} x_1 \\ \vdots \\ x_p \end{pmatrix}, \quad y = \begin{pmatrix} y_1 \\ \vdots \\ y_p \end{pmatrix}$$

$$d(x, y) = \sqrt{(x_1-y_1)^2 + \cdots + (x_p-y_p)^2} = \left(\sum (x_i - y_i)^2 \right)^{\frac{1}{2}}$$

(2) L^m distance

$$d(x,y) = \left(\sum_{i=1}^{p} |x_i - y_i|^m\right)^{\frac{1}{m}} \qquad m \geq 1$$

(3) Statistical distance

$$d(x,y) = \sqrt{(x-y)' S^{-1} (x-y)}$$

but then what's S?

(4) other distance

- Sometimes no meaningful p-dim measurement exists for items. One may introduce binary variables based on the presence and absence of a characteristic

1°) gender handedness

item i M(1) R(0) $\rightarrow$ $(1,0) = X_i$

item k F(0) L(1) $\rightarrow$ $(0,1) = X_k$

$$\underset{\substack{\text{item number} \quad j\text{th character}}}{X_{ij}} = \begin{cases} 1 & i\text{th item has } j\text{th character} \\ 0 & \text{no (otherwise)} \end{cases}$$

2°) dissimilarity

$$d^2(i,k) = d^2(X_i, X_k) = \sum_{j=1}^{p} (X_{ij} - X_{kj})^2$$

count of mismatches

- disadvantages:

Weighting 1-1 and 0-0 matches equally

- Similarity coefficients

1°) Contingency table

		item k		
item i		1	0	
	1	a	b	a+b
	0	c	d	c+d
		a+c	b+d	p = a+b+c+d

2°) Similarity coefficients

$d(i,k) = :$

(1) $\frac{a+d}{p} = \frac{a+d}{a+b+c+d}$

(2) $\frac{2(a+d)}{2(a+d)+c+b}$

(3) $\frac{a+d}{a+d+2(c+b)}$

(4) $\frac{a}{p}$

(5) $\frac{a}{a+b+c}$

(6) . . . ~

(C) similarity measures for pairs of variables

x_i, x_K

① Correlation Coefficient

② similarity coefficients for binary variables

- Contingency table

		variable K			
variable i		1	0		
	1	a	b	a+b	
	0	c	d	c+d	
		a+c	b+d	n=a+b+c+d	

	X_i	...	X_K
n obs	1		0
	0		1
	1	...	1
	0		1
	⋮		⋮

- Similarity Coefficients

$d(i,K) :=$

① $r_{ij} = \dfrac{\Sigma(X_{ij} - \bar{X}_i)(X_{Kj} - \bar{X}_K)}{\sqrt{\Sigma(X_{ij}-\bar{X}_i)^2}\sqrt{\Sigma(X_{Kj}-\bar{X}_K)^2}} = \dfrac{ad-bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$

② $\dfrac{a+d}{n}$

③ $\dfrac{2(a+d)}{2(a+d)+c+b}$

(d) Matrix D (Symmetric)

		item (variable				
		1	2	3	4	~
item (variable	1	d(1,1)			d(1,4)	
	2	d(2,1)	d(2,2)		d(2,4)	—
	⋮		⋮			

- Hierarchical clustering methods

(a) agglomerative and divisive hiherarchical methods

<1> agglomerate

Initially, each object is a cluster, then a series of successive mergers

(2) Divisive

Initially, only one cluster that contains all objects., then a series of successive divisions.

(b) Agglomerative hierarchical procedures

<1> clustering algorithm for grouping N objects.

- Input: A matrix $D = \{d_{ij}\}_{N \times N}$

  (distance. similarity, coordinates]

- Output: A tree diagran (dendrogram)
- steps

1°) start with N clusters. and a matrix $D = \{d_{ij}\}$

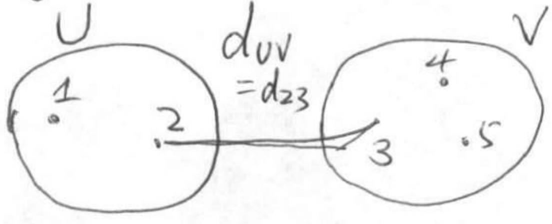
2°) Search D for the "most similar" (nearest) pair of clusters U and V

3° Merge U and V, labeled the new cluster UV. update D

4° Repeat 2°) and 3°) $N-1$ times

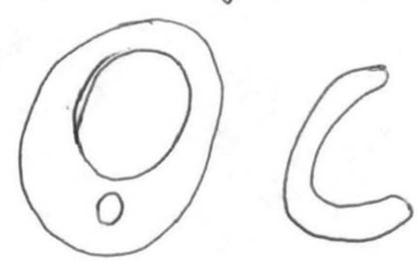
(2) Single linkage (method)

- Distance between two clusters is identified to be the smallest distance
- Merge nearest neighbors (pairs with the smallest distance or largest similarity)



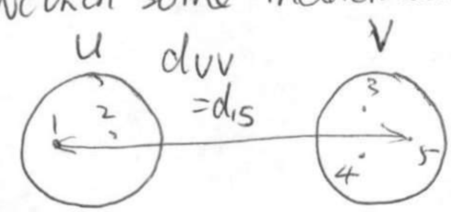
$$d_{UV} = \min_{\substack{u \in U \\ v \in V}} d_{uv}$$

- Can't discern (recognized, distinct, detect) poorly seperated clusters but good for un elliptical ones



(3) Complete linkage

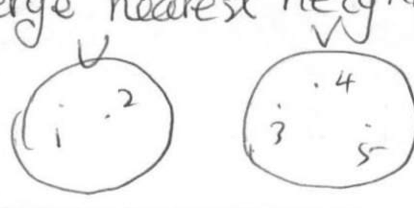
- Distance between two clusters is the maximum distance
- Merge nearest neighbors (so that all items in a cluster are within some maximum distance (or minimum similarity)



$$d_{UV} = \max_{\substack{u \in U \\ v \in V}} d_{uv}$$

(4) Average linkage

- Distance between two clusters is the average distance
- Merge nearest neighbors



$$d_{UV} = \frac{\sum_{i=1}^{2} \sum_{j=1}^{3} d_{ij}}{6} = \frac{\sum_{u \in U} \sum_{v \in V} d_{uv}}{N_u N_V}$$

(5) Ward's Hierarchical cluster's method

- Matrix $D = \{d_{ij}\}$ is coordinate matrix

$$
\begin{array}{cc|cccccc}
 & & & & \text{variables} & & & \\
 & & 1 & 2 & 3 & - & - & p \\
\hline
 & 1 & X_{11} & X_{12} & - & & - & X_{1p} \\
\text{items} & 2 & & & & & & \\
 & \vdots & \vdots & & & & & \\
 & N & X_{N1} & - & & - & & X_{Np}
\end{array}
$$

- Error sum of squares (ESS)

$$ESS_K = \sum_{X \in K} (X_i - \bar{X}_K)'(X_i - \bar{X}_K) \qquad K = 1 \sim g$$

g: # of groups

$$ESS = \sum_K ESS_K$$

- Initially, N clusters

$$ESS_K = 0, \quad K = 1 \cdots N \quad ESS = 0$$

- Merge pairs of clusters U an V such that the resulting increasing in ESS is minimum
- Good for observations that are roughly elliptically shaped.

(C) Examples

<1>

$$D = \begin{array}{c|ccccc} & 1 & 2 & 3 & 4 & 5 \\ \hline 1 & 0 & & & & \\ 2 & 9 & 0 & & & \\ 3 & 3 & 7 & 0 & & \\ 4 & 6 & 5 & 9 & 0 & \\ 5 & 11 & 10 & 2 & 8 & 0 \end{array}$$

1° Single linkage

- merge 3 and 5 → (3,5), 1, 2, 4
- update D

1°) Single linkage

$$\begin{array}{c|cccc} & (3,5) & 1 & 2 & 4 \\ \hline (3,5) & 0 & & & \\ 1 & 3 & 0 & & \\ 2 & 7 & 9 & 0 & \\ 4 & 8 & 6 & 5 & 0 \end{array}$$

1°) Single linkage

→ (1 3 5), 2, 4

2° complete linkage

$$\begin{array}{c|cccc} & (3,5) & 1 & 2 & 4 \\ \hline (3,5) & 0 & & & \\ 1 & 11 & 0 & & \\ 2 & 10 & 9 & 0 & \\ 4 & 9 & 6 & 5 & 0 \end{array}$$

→ (3,5) (2,4), 1

3° Average linkage

$$\begin{array}{c|cccc} & (3,5) & 1 & 2 & 4 \\ \hline (3,5) & 0 & & & \\ 1 & 7 & 0 & & \\ 2 & 8.5 & 9 & 0 & \\ 4 & 8.5 & 6 & 5 & 0 \end{array}$$

→ (3,5), (2,4), 1

12.4 Nonhierarchical clustering method

(a) objective and advantage

<1> objective

- cluster items (not variabbs) into K disjoint groups
- K is speified in advance or determined during clustering process

(2) advantage

- Good for large data set

  SAS cluster Fastclus

(b) K-means method

<1> algorithm

- Input: A coordinate matrix and a number K
- Output: K (disjoint) clusters
- step:

1°) partition items into K initial clusters.
Calculate the centroid (mean) for each cluster

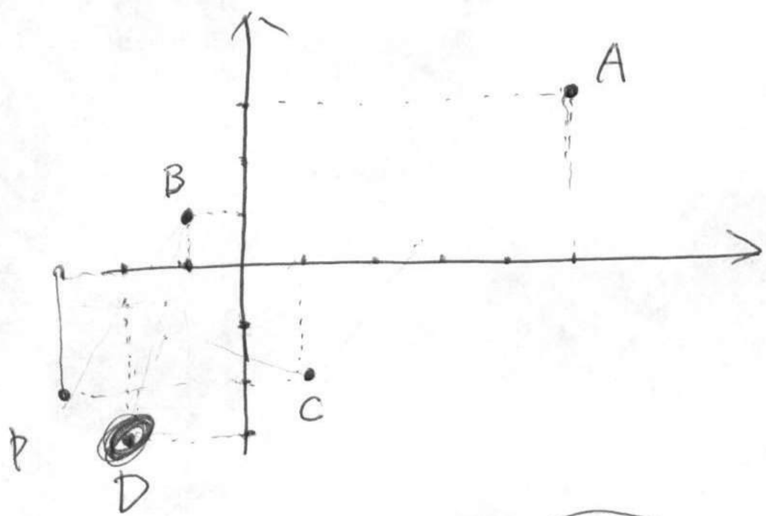
2°) Reassign an item to the cluster whose centroid (mean) is nearest to the item.

3°) Repeet 2°) untill no more reassignment take place

(2) Example

		Variables	
		X_1	X_2
items	A	5	3
	B	-1	1
	C	1	-2
	D	-3	-2

- objective: group items into 2 clusters



- start with AB and CD

(distance)^2

	centroid	A	B	C	D
AB	(2, 2)	10	10	17	41
CD	(-1, -2)	$6^2+5^2 = 61$	9	4	4

- end up ~~with~~ assigning B to CD $\Rightarrow$ A BCD

	centroid	A	B	C	D
A	5 3	0	$6^2+2^2=40$	41	89
BCD	(-1 -1)	$6^2+4^2=52$	4	5	5

- no more assigning, end up with A BCD

- However. if we start with (AC) and (BD)

  end up with (AC) and (BD)

- stat with B (ACD) $\longrightarrow$ (AC) (BD)

(C) Coments

(1) choice of K (no clear clud)

- Subject-matter knowledge
- data-based appraisals
- specify
- try different K's

(2) For incommensurable responses
standardize observations

(3) K-mean method is not robust against outliers

- Should try different distance measures

  L^m-distance $(0 < m < 1$

- Use multivariate median as centroid

  $\arg\min \sum \| x_i - \theta \|$

- Use multivariate weighted mean